

# Estimating the Upcrossings Index

J.R. Sebastião \*

Escola Superior de Gestão

Instituto Politécnico de Castelo Branco

Portugal

A.P. Martins, H. Ferreira, L. Pereira †

Departamento de Matemática

Universidade da Beira Interior

Portugal

**Abstract:** For stationary sequences, under general local and asymptotic dependence restrictions, any limiting point process for time normalized upcrossings of high levels is a compound Poisson process, i.e., there is a clustering of high upcrossings, where the underlying Poisson points represent cluster positions, and the multiplicities correspond to cluster sizes. For such classes of stationary sequences there exists the upcrossings index  $\eta$ ,  $0 \leq \eta \leq 1$ , which is directly related to the extremal index  $\theta$ ,  $0 \leq \theta \leq 1$ , for suitable high levels. In this paper we consider the problem of estimating the upcrossings index  $\eta$  for a class of stationary sequences satisfying a mild oscillation restriction. For the proposed estimator, properties such as consistency and asymptotic normality are studied. Finally, the performance of the estimator is assessed through simulation studies for autoregressive processes and case studies in the fields of environment and finance.

**Keywords:** Upcrossings index, local dependence conditions, consistency and asymptotic normality.

**Mathematics Subject Classification (2000)** 60G70

## 1 Introduction and preliminary results

Let  $\mathbf{X} = \{X_n\}_{n \geq 1}$  be a stationary sequence and  $\mathbf{u} = \{u_n\}_{n \geq 1}$  a sequence of real levels. We say that  $\mathbf{X}$  has an upcrossing of  $u_n$ , at  $i$ , if the event  $\{X_i \leq u_n < X_{i+1}\}$ ,  $i = 1, \dots, n$ , occurs. The point process of upcrossings of  $u_n$  by the first  $n$  variables of  $\mathbf{X}$  is then defined by

$$\tilde{N}_n(B) = \sum_{i=1}^{n-1} \mathbb{I}_{\{X_i \leq u_n < X_{i+1}\}} \delta_{\frac{i}{n}}(B), \quad B \subset [0, 1], \quad n \geq 1, \quad (1.1)$$

where  $\delta_a(\cdot)$  denotes unit mass at  $a \in B$  measure and  $\mathbb{I}_A$  the indicator of event  $A$ .

Ferreira [4] showed that if  $\mathbf{X}$  satisfies the mixing  $\Delta(\mathbf{u})$  condition, introduced in Hsing *et al.* [9], and  $\tilde{N}_n$  converges in distribution (as a point process on  $[0,1]$ ), then the limit is necessarily a compound Poisson process. For independent and identically distributed (i.i.d.) sequences it is easy to see that  $\tilde{N}_n$

---

\*Escola Superior de Gestão, Instituto Politécnico de Castelo Branco, 6000 Castelo Branco, Portugal.

E-mail:jrenato@esg.ipcb.pt

†Departamento de Matemática, Universidade da Beira Interior, 6200 Covilhã, Portugal. E-mail:amartins@mat.ubi.pt; lpereira@mat.ubi.pt; helena@mat.ubi.pt

converges in distribution if and only if  $u_n \equiv u_n^{(\nu)}$  for some  $\nu > 0$ , where  $u_n^{(\nu)}$  are normalized levels for upcrossings, that is,  $\mathbf{u}^{(\nu)} = \{u_n^{(\nu)}\}_{n \geq 1}$  denotes a sequence that satisfies

$$\lim_{n \rightarrow +\infty} nP(X_1 \leq u_n^{(\nu)} < X_2) = \nu, \quad (1.2)$$

being the limit point process a Poisson process on  $[0,1]$  with intensity  $\nu$ . We remark that when  $\mathbf{X}$  is a stationary sequence, satisfying the long range dependence condition  $D(\mathbf{u}^{(\nu)})$  of Leadbetter [12] and the mild oscillation restriction  $D''(\mathbf{u}^{(\nu)})$  of Leadbetter and Nandagopalan [13],  $\tilde{N}_n$  converges in distribution to the same Poisson process as in the i.i.d. case (Leadbetter and Nandagopalan [13]). When  $\mathbf{X}$  only satisfies condition  $\Delta(\mathbf{u}^{(\nu)})$  the limit compound Poisson process  $\tilde{N}$  is in fact a direct consequence of the clustering of high level upcrossings in a dependent sequence. Its intensity is simply  $\eta\nu$ , where  $\eta$  is the upcrossings index, and has multiplicity distribution  $\tilde{\pi}(j) = \lim_{n \rightarrow +\infty} \tilde{\pi}_n(j)$ ,  $j = 1, 2, \dots$ , with

$$\tilde{\pi}_n(j) = P\left(\sum_{i=1}^{r_n} \mathbb{I}_{\{X_i \leq u_n < X_{i+1}\}} = j \mid \sum_{i=1}^{r_n} \mathbb{I}_{\{X_i \leq u_n < X_{i+1}\}} > 0\right), \quad j = 1, 2, \dots \quad (1.3)$$

for some sequence  $r_n = \lfloor n/k_n \rfloor$  and  $\{k_n\}_{n \geq 1}$  a sequence of positive integers satisfying

$$k_n \xrightarrow{n \rightarrow +\infty} +\infty, \quad \frac{k_n l_n}{n} \xrightarrow{n \rightarrow +\infty} 0, \quad k_n \alpha_{n, l_n} \xrightarrow{n \rightarrow +\infty} 0, \quad (1.4)$$

where  $\alpha_{n, l_n}$  are the mixing coefficients of the  $\Delta(\mathbf{u})$  condition.

The upcrossings of  $u_n$  by  $X_i$  with  $i \in J_{n,j} = \{(j-1)r_n + 1, \dots, jr_n\}$ , for some  $j = 1, \dots, k_n$ , are regarded as forming a cluster of upcrossings and  $\tilde{\pi}_n(j)$ ,  $j = 1, 2, \dots$  is called the conditional cluster size distribution, since it is simply the distribution of the number of upcrossings in a cluster (i.e. in a set  $J_{n,j}$ ) given that there is at least one.

The upcrossings index can then be viewed as a measure of clustering of upcrossings of high levels  $u_n$  by the variables in  $\mathbf{X}$  and is formally defined as follows.

**Definition 1.1 (Ferreira [4])** *If for each  $\nu > 0$  there exists  $\mathbf{u}^{(\nu)} = \{u_n^{(\nu)}\}_{n \geq 1}$  and  $\lim_{n \rightarrow +\infty} P(\tilde{N}_n^{(\nu)}([0,1]) = 0) = e^{-\eta\nu}$ , for some constant  $0 \leq \eta \leq 1$ , then we say that the sequence  $\mathbf{X}$  has upcrossings index  $\eta$ .*

Many common cases, as i.i.d. sequences or sequences that satisfy condition  $D''(\mathbf{u}^{(\nu)})$ , have upcrossings index  $\eta = 1$ . A value  $\eta < 1$  indicates clustering of upcrossings of  $\mathbf{u}^{(\nu)}$ , giving rise to multiplicities in the limit.

If for each  $\nu > 0$  there exists  $u_n^{(\nu)}$  satisfying (1.2) and  $\lim_{n \rightarrow +\infty} P(X_1 > u_n^{(\nu)}) = \tau$  for some  $\tau > 0$  then the upcrossings index  $\eta$  exists if and only if the extremal index  $\theta$  exists and, in this case,

$$\theta = \frac{\nu}{\tau} \eta. \quad (1.5)$$

(Ferreira [4]). Note that the previous conditions imply that  $\theta \leq \eta$  since if a level  $u_n$  is simultaneously normalized for upcrossings and for exceedances we necessarily have  $\nu \leq \tau$ .

The conditional cluster size distribution  $\tilde{\pi}_n$  defined in (1.3) is also related to the upcrossings index  $\eta$  in the following interesting manner, which is a direct consequence of Lemma 2.1 of Ferreira [4].

**Proposition 1.1** Suppose  $\mathbf{X}$  satisfies  $\Delta(\mathbf{u}^{(\nu)})$ ,  $\nu > 0$ , has upcrossings index  $\eta > 0$  and let  $\tilde{\pi}_n$  be defined as in (1.3) with  $u_n = u_n^{(\nu)}$ . Then

$$\lim_{n \rightarrow +\infty} \sum_{j \geq 1} j \tilde{\pi}_n(j) = \frac{1}{\eta}.$$

The reciprocal of the upcrossings index can then be interpreted as the limiting mean cluster size of upcrossings. Hence, it is natural to estimate  $\eta$  as the reciprocal of the sample average cluster size. However, such an estimator suffers from the drawback that the identification of clusters (equivalently, choice of  $k_n$ ) depends on the knowledge of the mixing coefficients  $\alpha_{n,l_n}$ . Alternative ways of identifying clusters of high level upcrossings is therefore a key issue for the estimation of  $\eta$ .

If a sequence is suitably well-behaved then one might hope that groups of successive upcrossings of  $u_n$  are sufficiently far apart that each group can be regarded as a separate cluster. A sufficient condition for this to hold is condition  $\tilde{D}^{(3)}(\mathbf{u})$  of Ferreira [4].

**Definition 1.2 (Ferreira [4])** Let  $\mathbf{X}$  be a sequence satisfying the condition  $\Delta(\mathbf{u})$ .  $\mathbf{X}$  satisfies condition  $\tilde{D}^{(3)}(\mathbf{u})$  if

$$\lim_{n \rightarrow +\infty} nP(X_1 \leq u_n < X_2, \tilde{N}_{3,3} = 0, \tilde{N}_{4,r_n} > 0) = 0,$$

for some sequence  $r_n = [n/k_n]$  with  $\{k_n\}_{n \geq 1}$  satisfying (1.4) and  $\tilde{N}_{i,j} \equiv \tilde{N}_n([i/n, j/n])$  with  $\tilde{N}_{i,j} = 0$  if  $j < i$ .

**Remark 1** Condition  $\tilde{D}^{(3)}(\mathbf{u})$  is implied by condition

$$\lim_{n \rightarrow \infty} n \sum_{j=4}^{r_n-1} P(X_1 \leq u_n < X_2, \tilde{N}_{3,3} = 0, X_j \leq u_n < X_{j+1}) = 0,$$

which is clearly implied by condition

$$\lim_{n \rightarrow \infty} n \sum_{j=4}^{r_n-1} P(X_1 \leq u_n < X_2, X_j \leq u_n < X_{j+1}) = 0. \quad (1.6)$$

**Remark 2** Sequences that satisfy condition  $D''(\mathbf{u})$  also satisfy condition  $\tilde{D}^{(3)}(\mathbf{u})$  and the latter belongs to a wide family of local dependence conditions  $\tilde{D}^{(k)}(\mathbf{u})$ ,  $k \geq 2$ , introduced in Ferreira [4]. This family of local dependence conditions is slightly stronger than the family of conditions  $D^{(k)}(\mathbf{u})$ ,  $k \geq 1$ , considered in Chernick et al. [2]. The negatively correlated uniform AR(1) process given in Chernick et al. [2] is an example of a process that verifies the previous condition as shown in Sebastião et al. [18]. In Ferreira [4] we find an example of a max-autoregressive process for which  $\tilde{D}^{(3)}(\mathbf{u})$  also holds.

**Remark 3** Necessary and sufficient conditions for (1.6) to hold, involving upcrossings-tail dependence coefficients, can be found in Ferreira and Ferreira [6].

Condition  $\tilde{D}^{(3)}(\mathbf{u})$  locally restricts the dependence of the sequence, but still allows clustering of upcrossings. It roughly states that whenever an upcrossing of a high level occurs, a cluster of upcrossings may follow it, but once the sequence falls below the threshold it is very unlikely to upcross it again in the nearby observations. Thus, it enables the identification of upcrossings clusters since for this class

upcrossings may be simply identified asymptotically as runs of consecutive upcrossings and the cluster sizes as run lengths. Indeed, Ferreira [5] proved that if the conditional upcrossing run length distribution is defined as

$$\tilde{\pi}_n^*(j) = P(X_3 \leq u_n < X_4, \dots, X_{2j+1} \leq u_n < X_{2j+2}, \tilde{N}_{2j+3,2j+3} = 0 \mid \tilde{N}_{1,1} = 0, X_3 \leq u_n < X_4), \quad j \geq 1, \quad (1.7)$$

then the following result gives the conditional expected length of an upcrossing run.

**Proposition 1.2 (Ferreira [5])** *If, for each  $\nu > 0$ ,  $\mathbf{X}$  satisfies condition  $\tilde{D}^{(3)}(\mathbf{u}^{(\nu)})$  and  $P(X_3 \leq u_n^{(\nu)} < X_4, \dots, X_{2j-1} \leq u_n^{(\nu)} < X_{2j}) \rightarrow 0$ , as  $j \rightarrow +\infty$ , then the upcrossings index of  $\mathbf{X}$  exists and is equal to  $\eta$  if and only if,*

$$\sum_{j \geq 1} j \tilde{\pi}_n^*(j) = \frac{P(X_1 \leq u_n^{(\nu)} < X_2)}{P(X_1 \leq u_n^{(\nu)} < X_2, \tilde{N}_{3,3} = 0)} \xrightarrow{n \rightarrow +\infty} \frac{1}{\eta}, \quad (1.8)$$

for each  $\nu > 0$ .

From this result it seems natural to estimate  $\eta$  by the reciprocal of the sample average run length. We shall consider such an estimator in Section 3.

**Remark 4** *Under condition  $\tilde{D}^{(3)}(\mathbf{u})$  the upcrossings index can be related to other dependence measures, namely the lag-1 upcrossings tail dependence coefficient  $\mu_1$ , since  $\eta = 1 - \mu_1 = 1 - \lim_{x \rightarrow x_F} P(X_3 \leq u_n^{(\nu)} < X_4 \mid X_1 \leq u_n^{(\nu)} < X_2)$ , where  $x_F$  is the upper limit of the common marginal distribution  $F$  of  $\mathbf{X}$ , as shown in Ferreira and Ferreira [6].*

Alternative expressions for  $\eta$  involving only stationarity are given in the following simple lemma.

**Lemma 1.1** *If  $nP(X_1 \leq u_n < X_2, \tilde{N}_{3,3} = 0) \xrightarrow{n \rightarrow +\infty} \xi > 0$  then the following are equivalent:*

- i)  $P(\tilde{N}_{3,3} = 0 \mid X_1 \leq u_n < X_2) \xrightarrow{n \rightarrow +\infty} \eta$ ;
- ii)  $nP(X_1 \leq u_n < X_2) \xrightarrow{n \rightarrow +\infty} \frac{\xi}{\eta}$ ;
- iii)  $n(1 - P(\tilde{N}_{1,1} = 0, \tilde{N}_{3,3} = 0)) \xrightarrow{n \rightarrow +\infty} \xi + \frac{\xi}{\eta}$ .

**Remark 5** *From the stationarity of  $\mathbf{X}$  we obviously have  $P(X_1 \leq u_n < X_2, \tilde{N}_{3,3} = 0) = P(\tilde{N}_{1,1} = 0, X_3 \leq u_n < X_4)$  and consequently an upcrossings run can be either identified at its beginning or at its end.*

The upcrossings index estimation is important not only by itself, as a measure of clustering of upcrossings of high levels, but also because of its relation with other dependence coefficients, namely the extremal index and the lag-1 upcrossings tail coefficient (Ferreira and Ferreira [6]). Therefore, we shall pursue this issue in the remainder of this paper. Note that under the conditions of Proposition 1.2 it is reasonable to estimate  $\eta$  by the ratio between the number of upcrossings followed by a non-upcrossing and the number of upcrossings of a high level.

In Section 2, we formally present such an estimator, suggested in Ferreira [5], which we shall call the runs estimator of the upcrossings. In the subsequent sections we show that such an estimator is

typically weakly consistent and asymptotically normal. In Section 6 we carry out a simulation study of finite sample behaviour of the proposed estimator in a max-autoregressive process and in a first order autoregressive process. In Section 7 the performance of the estimator is also assessed through case studies in the fields of environment and finance. Conclusions are found in Section 8.

## 2 The runs estimator of the upcrossings index

Before formally defining the runs estimator of the upcrossings index, we shall start by introducing some notation that will be used throughout the paper.

Let  $\widehat{N}_n$  denote the point process of non-upcrossings of  $u_n$ , followed by an upcrossing, by the first  $n$  variables of  $\mathbf{X}$ , that is

$$\widehat{N}_n(B) = \sum_{i=1}^{n-3} \mathbb{I}_{\{\widetilde{N}_{i,i}=0, X_{i+2} \leq u_n < X_{i+3}\}} \delta_{\frac{i}{n}}(B), \quad B \subset [0, 1], \quad n \geq 1. \quad (2.1)$$

For a sequence of real levels  $\{u_n\}_{n \geq 1}$  let's define the random variables  $Y_i \equiv Y_i(u_n)$ ,  $i = 1, \dots, n$ , corresponding to the number of consecutive upcrossings of the level  $u_n$  occurring from instant  $i + 2$  on, that is,

$$Y_i = \begin{cases} 0 & \text{if } \widetilde{N}_{i+2,i+2} = 0 \\ k & \text{if } X_{i+2} \leq u_n < X_{i+3}, X_{i+4} \leq u_n < X_{i+5}, \dots, X_{i+2k} \leq u_n < X_{i+2k+1}, \widetilde{N}_{i+2k+2,i+2k+2} = 0, \end{cases}$$

with  $k \geq 1$ . Furthermore, let's denote by  $Z_i(u_n)$ ,  $i = 1, \dots, n$ , the length of each of these sequences given the occurrence of a non-upcrossing followed by an upcrossing at instant  $i + 2$ , of level  $u_n$ , which has distribution  $\widetilde{\pi}_n^*(\cdot)$  given in (1.3) since

$$\widetilde{\pi}_n^*(k) = P(Z_i(u_n) = k) = P(Y_i = k \mid \widetilde{N}_{i,i} = 0, X_{i+2} \leq u_n < X_{i+3}), \quad k \geq 1, \quad (2.2)$$

independent of  $i$  from the stationarity of  $\mathbf{X}$ .

If  $\mathbf{X}$  has upcrossings index  $\eta > 0$  and the conditions of Proposition 1.2 hold then for  $u_n = u_n^{(\nu)}$

$$\begin{aligned} \frac{1}{\eta} &= \lim_{n \rightarrow +\infty} E[Z_1(u_n)] = \lim_{n \rightarrow +\infty} E[Y_1 \mid \widetilde{N}_{1,1} = 0, X_3 \leq u_n < X_4] \\ &= \lim_{n \rightarrow +\infty} \frac{P(X_1 \leq u_n < X_2)}{P(\widetilde{N}_{1,1} = 0, X_3 \leq u_n < X_4)} = \lim_{n \rightarrow +\infty} \frac{E[\widetilde{N}_n([0, 1])]}{E[\widehat{N}_n([0, 1])]}. \end{aligned}$$

From this result it is natural to propose the non parametric estimator for  $\eta$  given by the ratio between the total number of non-upcrossings followed by an upcrossings and the total number of upcrossings

$$\widehat{\eta}_n = \widehat{\eta}_n(u) := \frac{\widehat{N}_n([0, 1])}{\widetilde{N}_n([0, 1])} = \frac{\sum_{i=1}^{n-3} \mathbb{I}_{\{\widetilde{N}_{i,i}=0, X_{i+2} \leq u < X_{i+3}\}}}{\sum_{i=1}^{n-1} \mathbb{I}_{\{X_i \leq u < X_{i+1}\}}}, \quad (2.3)$$

where  $u$  is a suitable threshold.

We shall call this estimator the runs estimator of the upcrossings index attending to its similarity with the runs estimator of the extremal index proposed by Nandagopalan [15] for stationary sequences satisfying condition  $D''(\mathbf{u})$ , as suggested in Ferreira [5]. In practical applications we will always use this

estimator (2.3), nevertheless, the theoretical properties presented in the following sections will be proved for an estimator that is asymptotically equivalent to  $\hat{\eta}_n$  and that we define in what follows.

Let  $\hat{N}_n^{(T)}$  denote the marked point process on  $[0, 1]$  defined by

$$\hat{N}_n^{(T)}(B) = \sum_{i=1}^{n-3} T(Y_i) \mathbb{I}_{\{\tilde{N}_{i,i}=0, X_{i+2} \leq u_n < X_{i+3}\}} \delta_{\frac{i}{n}}(B), \quad B \subset [0, 1], \quad n \geq 1, \quad (2.4)$$

where  $T : \mathbb{N} \rightarrow \mathbb{R}$  is a given mapping. Thus,  $\hat{N}_n^{(T)}$  has mass equal to  $T(Y_i)$  at the point  $i/n$  whenever  $\mathbf{X}$  has a non-upcrossing at  $i$  followed by an upcrossing.

If in (2.4) we consider  $T(y) \equiv 1$  we obtain the point process  $\hat{N}_n$  in (2.1). Whereas, if we consider  $T(y) = y \geq 1$  we obtain a point process, that we shall denote by  $\bar{N}_n$ , which differs from the upcrossings point process  $\tilde{N}_n$  in (1.1) on an event with probability bounded by  $P(X_1 \leq u_n < X_2)$ , which converges to zero, as  $n \rightarrow +\infty$ , since the levels  $u_n$  are commonly chosen to satisfy (1.2). Therefore, we can instead consider the estimator  $\hat{\eta}_n^* = \hat{N}_n([0, 1]) / \bar{N}_n([0, 1])$  since, for such levels,  $\hat{\eta}_n = \hat{\eta}_n^* + o_p(1)$ . The properties proved, in what follows, for  $\hat{\eta}_n^*$ , then also apply to  $\hat{\eta}_n$  that we shall use in the simulation studies.

### 3 Weak consistency

We show, in this section, that  $\hat{\eta}_n^*$  is a weakly consistent estimator of  $\eta$  under mild assumptions.

Throughout it will be assumed that  $\mathbf{X}$  is a stationary sequence satisfying condition  $\tilde{D}^{(3)}(\mathbf{u})$  and has upcrossings index  $\eta > 0$ .

Note that when the level  $u_n$  satisfies (1.2) there are insufficient upcrossings to give statistical “consistency” for the estimator  $\hat{\eta}_n^*$ . That is, as  $n$  increases the value of  $\hat{\eta}_n^*$  does not necessarily converge appropriately to the value  $\eta$ . Nevertheless, consistency can be achieved by the use of somewhat lower levels. We shall therefore consider non-normalized levels, in the sense of (1.2),  $v_n = u_{[n/c_n]}^{(\nu)}$ , for some fixed  $\nu > 0$ , that satisfy

$$nP(X_1 \leq v_n < X_2) - c_n \nu \xrightarrow{n \rightarrow +\infty} 0 \quad (3.1)$$

where  $\{c_n\}_{n \geq 1}$  and  $\{k_n\}_{n \geq 1}$  are sequences of real levels such that  $c_n, k_n \xrightarrow{n \rightarrow +\infty} +\infty$  and  $c_n/k_n \xrightarrow{n \rightarrow +\infty} 0$ . Note that for this sequence of levels we also have  $\lim_{n \rightarrow +\infty} E[Z_1(v_n)] = 1/\eta$ .

With the same arguments used in Nandagopalan [15] we obtain the following lemma which is essential to obtain the properties of the estimator  $\hat{\eta}_n^*$  that we further present.

**Lemma 3.1** *Suppose  $\{k_n\}_{n \geq 1}$  is a sequence of integers such that  $k_n \xrightarrow{n \rightarrow +\infty} +\infty$  and that there exists a sequence  $\{l_n\}_{n \geq 1}$  for which*

$$k_n[\alpha_{n, l_n-2} + P(\tilde{N}_n([0, l_n/n]) > 0)] \xrightarrow{n \rightarrow +\infty} 0. \quad (3.2)$$

Then

$$E[e^{ia_n \hat{N}_n^{(T)}}(J_n)] - \prod_{j=1}^{k_n} E[e^{ia_n \hat{N}_n^{(T)}}(J_{nj})] \xrightarrow{n \rightarrow +\infty} 0,$$

for any sequence of real numbers  $\{a_n\}_{n \geq 1}$ , where  $J_n \subset [0, 1]$ ,  $n \geq 1$ , is a sequence of intervals such that for each  $n$ ,  $J_n \supset \bigcup_{j=1}^{k_n} J_{nj}$ , with  $J_{nj}$ ,  $j = 1, \dots, k_n$  disjoint subintervals satisfying  $m(J_n) - m(\bigcup_{j=1}^{k_n} J_{nj}) \leq k_n/n$  ( $m(\cdot)$  denoting Lebesgue measure).

**Remark 6** Condition (3.2) is satisfied for any sequence of normalized levels  $\mathbf{u}^{(\nu)}$ , as in (1.2), if condition  $\Delta(\mathbf{u}^{(\nu)})$  holds, since this condition implies that  $k_n \alpha_{n, l_n-2} \xrightarrow{n \rightarrow +\infty} 0$  and, for levels  $\mathbf{u}^{(\nu)}$ ,  $k_n P(\tilde{N}_n([0, l_n/n]) > 0) \xrightarrow{n \rightarrow +\infty} 0$ .

Henceforth, we shall write  $\overline{N}_n = \overline{N}_n([0, 1]) = \sum_{i=1}^{n-3} Y_i \mathbb{I}_{\{N_{i,i}=0, X_{i+2} \leq u_n < X_{i+3}\}}$  and  $\hat{N}_n = \hat{N}_n([0, 1]) = \sum_{i=1}^{n-3} \mathbb{I}_{\{N_{i,i}=0, X_{i+2} \leq u_n < X_{i+3}\}}$  so that  $\overline{N}_n$  and  $\hat{N}_n$  will now denote random variables rather than point processes, and accordingly  $\overline{N}_{r_n} = \overline{N}_n([0, r_n/n])$  and  $\hat{N}_{r_n} = \hat{N}_n([0, r_n/n])$ .

Lets consider the compound Poisson random variable

$$\overline{N}_n^* = \sum_{j=1}^{\hat{N}_n^*} Z_j^*(v_n)$$

where  $\hat{N}_n^*$  denotes a Poisson random variable with mean  $c_n \eta \nu$  and the random variables  $Z_j^*(v_n)$  are independent and identically distributed with the same distribution as  $Z_1(v_n)$ , that is, distribution  $\pi_n^*$  given in (2.2).

In the following results we prove, with similar arguments to the ones used in Nandagopalan [15], that under suitable conditions the limiting distributions of  $\overline{N}_n$  and  $\hat{N}_n$  are identical to, respectively, those of  $\overline{N}_n^*$  and  $\hat{N}_n^*$ .

**Theorem 3.1** Let  $\{v_n\}_{n \geq 1}$  be a sequence of levels satisfying (3.1) and suppose there exists a sequence  $\{l_n\}_{n \geq 1}$  for which (3.2) holds for such levels.

If

$$E \left[ Z_j(v_n) \mathbb{I}_{\{\hat{N}_{r_n} > 1\}} \right] = o(1) \quad (3.3)$$

uniformly in  $j = 1, \dots, r_n$ , ( $r_n = \lfloor n/k_n \rfloor$ ), then

$$E \left[ \exp(itc_n^{-1} \overline{N}_n) \right] - E \left[ \exp(itc_n^{-1} \overline{N}_n^*) \right] \xrightarrow{n \rightarrow +\infty} 0 \quad (3.4)$$

and

$$E \left[ \exp(itc_n^{-1} \hat{N}_n) \right] - E \left[ \exp(itc_n^{-1} \hat{N}_n^*) \right] \xrightarrow{n \rightarrow +\infty} 0, \quad (3.5)$$

for each  $t \in \mathbb{R}$ .

**Proof:** To prove (3.4), let  $\Psi_{Z_1}(t) = E[\exp(itZ_1)]$  denote the characteristic function of the random variable  $Z_1$  with distribution  $\tilde{\pi}_n^*(\cdot)$  given in (2.2). Since  $\overline{N}_n^*$  is a compound Poisson variable we have

$$E \left[ \exp(itc_n^{-1} \overline{N}_n^*) \right] = \exp(-c_n \nu \eta (1 - \Psi_{Z_1}(tc_n^{-1})))$$

and hence it suffices to show that

$$E \left[ \exp(itc_n^{-1} \overline{N}_n) \right] = \exp(-c_n \nu \eta (1 - \Psi_{Z_1}(tc_n^{-1}))) + o(1). \quad (3.6)$$

Nevertheless, from Lemma 3.1, we can write

$$E [\exp(itc_n^{-1}\overline{N}_n)] = (E [\exp(itc_n^{-1}\overline{N}_{r_n})])^{k_n} + o(1).$$

Now following the steps of the proof of Proposition 5.3.1 of Nandagopalan [15] we obtain

$$E [\exp(itc_n^{-1}\overline{N}_{r_n})] = 1 - \frac{c_n}{k_n}(\eta\nu + o(1)) \times (1 - \Psi_{Z_1}(tc_n^{-1})) + o(k_n^{-1}),$$

since (3.3) holds and  $\frac{r_n}{c_n}k_n P(N_{1,1} = 0, X_3 \leq v_n < X_4) = \eta\nu + o(1)$  as a consequence of condition  $\tilde{D}^{(3)}(v_n)$ . Moreover, since  $|1 - \Psi_{Z_1}(tc_n^{-1})| \leq |tc_n^{-1}|E[Z_1(v_n)]$  and  $E[Z_1(v_n)] \xrightarrow{n \rightarrow +\infty} \eta^{-1}$  we have

$$E [\exp(itc_n^{-1}\overline{N}_n)] = \left(1 - \frac{c_n}{k_n}\eta\nu(1 - \Psi_{Z_1}(tc_n^{-1})) + o(k_n^{-1})\right)^{k_n} + o(1),$$

from which results (3.6) and thus proves (3.4).

Convergence (3.5) can be established with similar arguments.  $\square$

**Remark 7** *If in the previous result we consider normalized levels  $u_n^{(\nu)}$  then both convergence (3.4) and (3.5) would hold without the need of the normalizing constant  $c_n$ , being, in this case,  $\hat{N}_n^*$  a Poisson random variable with mean  $\eta\nu$ .*

The next result justifies the need to consider lower levels  $v_n$  satisfying (3.1) in order to guarantee the consistency of the estimator.

**Theorem 3.2** *Suppose  $\{v_n\}_{n \geq 1}$  is a sequence of levels satisfying (3.1),*

$$E [Z_1^*(v_n) \mathbb{I}_{\{Z_1^*(v_n) > c_n\}}] \xrightarrow{n \rightarrow +\infty} 0 \quad (3.7)$$

and

$$E [(Z_1^*(v_n))^2 \mathbb{I}_{\{Z_1^*(v_n) \leq c_n\}}] \xrightarrow{n \rightarrow +\infty} 0, \quad (3.8)$$

then

$$c_n^{-1}\overline{N}_n^* \xrightarrow[n \rightarrow +\infty]{P} \nu.$$

**Proof:** Start by noting that convergence (3.7) along with the fact that  $E[Z_1(v_n)] \xrightarrow{n \rightarrow +\infty} \eta^{-1}$  implies that  $\eta\nu E[Z_1^*(v_n) \mathbb{I}_{\{Z_1^*(v_n) \leq c_n\}}] \xrightarrow{n \rightarrow +\infty} \nu$ . It is then sufficient to show that

$$c_n^{-1}\overline{N}_n^* - \eta\nu E[Z_1^*(v_n) \mathbb{I}_{\{Z_1^*(v_n) \leq c_n\}}] \xrightarrow[n \rightarrow +\infty]{P} 0.$$

Now evoking similar arguments to the ones used by Nandagopalan [15] (Proposition 5.3.2) and considering  $\gamma_n = E[Z_1^*(v_n) \mathbb{I}_{\{Z_1^*(v_n) \leq c_n\}}]$ , we obtain, for any  $\epsilon > 0$  and  $\delta < \epsilon / \max\{\gamma_n\}$ ,

$$\begin{aligned} & P(|c_n^{-1}\overline{N}_n^* - \eta\nu\gamma_n| > 2\epsilon) \\ & \leq (\delta + \eta\nu)E[Z_1^*(v_n) \mathbb{I}_{\{Z_1^*(v_n) > c_n\}}] + \frac{E[(Z_1^*(v_n) \mathbb{I}_{\{Z_1^*(v_n) \leq c_n\}})^2]}{c_n(\eta\nu - \delta) + (\epsilon/(\delta + \eta\nu))^2} + P(|c_n^{-1}\hat{N}_n^* - c_n\eta\nu| > \delta c_n). \end{aligned}$$

The first two terms tend to zero by (3.7) and (3.8) respectively, while the third term tends to zero since  $\hat{N}_n^*$  is a compound Poisson random variable with mean  $c_n\eta\nu$ . Hence, since  $\epsilon > 0$  is arbitrary, we obtain  $c_n^{-1}\overline{N}_n^* - \eta\nu\gamma_n \xrightarrow[n \rightarrow +\infty]{P} 0$  as required.  $\square$



We proved in Theorem 3.1 that  $c_n^{-1}\overline{N}_n$  and  $c_n^{-1}\widehat{N}_n$  have the same asymptotic distribution as, respectively,  $c_n^{-1}\overline{N}_n^*$  and  $c_n^{-1}\widehat{N}_n^*$ , and in Theorem 3.2 proved that  $c_n^{-1}\overline{N}_n^* \xrightarrow[n \rightarrow +\infty]{P} \nu$ . Moreover, since  $\widehat{N}_n^*$  is a Poisson random variable with mean  $c_n\eta\nu$  it follows that  $c_n^{-1}\widehat{N}_n^* \xrightarrow[n \rightarrow +\infty]{P} \eta\nu$ . The weak consistency of  $\widehat{\eta}_n^*$  is now an immediate consequence, summarized in the next result.

**Corollary 3.1** *If the conditions of Theorems 3.1 and 3.2 hold then*

$$\widehat{\eta}_n^* \xrightarrow[n \rightarrow +\infty]{P} \eta.$$

## 4 Asymptotic Normality

Imposing additional conditions on the limiting behaviour of the first and second moments of the variable  $Z_1(v_n)$  we obtain in this section the asymptotic normality of our estimator  $\widehat{\eta}_n^*$ .

The proof of the next result, which is essential in obtaining what follows, shall be omitted since it is similar to the proof of Proposition 5.4.1 of Nandagopalan [15].

**Theorem 4.1** *Suppose that  $\{v_n\}_{n \geq 1}$  is a sequence of levels satisfying (3.1) and suppose there exists a sequence  $\{l_n\}_{n \geq 1}$  for which (3.2) holds for such levels.*

*If*

$$E[\overline{N}_{r_n}^2 \mathbb{I}_{\{\widehat{N}_{r_n} > 1\}}] = o(c_n/k_n),$$

$$E[(Z_j(v_n))^2 \mathbb{I}_{\{\widehat{N}_{r_n} > 1\}}] = o(1) \tag{4.1}$$

*uniformly in  $j = 1, \dots, r_n$ , ( $r_n = \lfloor n/k_n \rfloor$ ) and*

$$\sigma_n^2 = E[(Z_1(v_n))^2]$$

*is a bounded sequence, then  $c_n^{-1} \begin{bmatrix} \overline{N}_n - E[\overline{N}_n] \\ \widehat{N}_n - E[\widehat{N}_n] \end{bmatrix}$  converges in distribution if and only if  $c_n^{-1} \begin{bmatrix} \overline{N}_n^* - E[\overline{N}_n^*] \\ \widehat{N}_n^* - E[\widehat{N}_n^*] \end{bmatrix}$  does and, in this case, the limits coincide.*

**Theorem 4.2** *Suppose that  $\{v_n\}_{n \geq 1}$  is a sequence of levels satisfying (3.1) and suppose there exists a sequence  $\{l_n\}_{n \geq 1}$  for which (3.2) holds for such levels.*

*If*

$$\sigma_n^2 = E[(Z_1(v_n))^2] \xrightarrow[n \rightarrow +\infty]{} \sigma^2 < +\infty \tag{4.2}$$

*and for each  $\epsilon > 0$ ,*

$$E[(Z_1^*(v_n))^2 \mathbb{I}_{\{(Z_1^*(v_n))^2 > \epsilon c_n\}}] \xrightarrow[n \rightarrow +\infty]{} 0. \tag{4.3}$$

*Then*

$$c_n^{-1/2} \begin{bmatrix} \overline{N}_n^* - E[\overline{N}_n^*] \\ \widehat{N}_n^* - E[\widehat{N}_n^*] \end{bmatrix} \xrightarrow[n \rightarrow +\infty]{d} \mathcal{N} \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \nu\eta\sigma^2 & \nu \\ \nu & \nu\eta \end{bmatrix} \right].$$

**Proof:** Since  $\overline{N}_n^*$  is a compound Poisson random variable it holds  $E[\overline{N}_n^*] = E[\hat{N}_n^*]E[Z_1(v_n)]$  and we can write

$$\overline{N}_n^* - E[\overline{N}_n^*] = (\overline{N}_n^* - \hat{N}_n^*E[Z_1(v_n)]) + E[Z_1(v_n)](\hat{N}_n^* - E[\hat{N}_n^*]). \quad (4.4)$$

From Lindeberg's condition (4.3) the central limit theorem holds, this is,

$$c_n^{-1/2}(\overline{N}_n^* - \hat{N}_n^*E[Z_1(v_n)]) \xrightarrow[n \rightarrow +\infty]{d} \mathcal{N}\left(0, \eta\nu\left(\sigma^2 - \frac{1}{\eta^2}\right)\right)$$

and therefore

$$\begin{aligned} E[\exp(itc_n^{-1/2}(\overline{N}_n^* - \hat{N}_n^*E[Z_1(v_n)]))] &= \exp(-c_n\eta\nu[1 - \Psi_{Z_1}(tc_n^{-1/2})e^{-itc_n^{-1/2}E[Z_1(v_n)]}]) \\ &\xrightarrow[n \rightarrow +\infty]{} \exp\left(\frac{\eta\nu}{2}\left(\sigma^2 - \frac{1}{\eta^2}\right)t^2\right). \end{aligned}$$

Also,  $c_n^{-1/2}(\hat{N}_n^* - E[\hat{N}_n^*]) \xrightarrow[n \rightarrow +\infty]{d} \mathcal{N}(0, \eta\nu)$ , so that

$$E[\exp\{isc_n^{-1/2}(\hat{N}_n^* - E[\hat{N}_n^*])\}] = \exp(-c_n\eta\nu(1 - e^{isc_n^{-1/2}} + isc_n^{-1/2})) \xrightarrow[n \rightarrow +\infty]{} \exp\left(-\frac{\eta\nu s^2}{2}\right).$$

Now

$$\begin{aligned} E[\exp\{itc_n^{-1/2}(\overline{N}_n^* - \hat{N}_n^*E[Z_1(v_n)]) + isc_n^{-1/2}(\hat{N}_n^* - E[\hat{N}_n^*])\}] \\ = \exp(-c_n\eta\nu(1 - \Psi_{Z_1}(tc_n^{-1/2})e^{-itc_n^{-1/2}E[Z_1(v_n)]}e^{isc_n^{-1/2}} + isc_n^{-1/2})), \end{aligned} \quad (4.5)$$

with  $\Psi_{Z_1}(tc_n^{-1/2}) = 1 + itc_n^{-1/2}E[Z_1(v_n)] + \rho_n^{(1)}$  and  $e^{-itc_n^{-1/2}E[Z_1(v_n)]} = 1 - itc_n^{-1/2}E[Z_1(v_n)] + \rho_n^{(2)}$  where  $\rho_n^{(1)}$  and  $\rho_n^{(2)}$  are bounded by  $t^2c_n^{-1}\sigma^2/2$ .

Thus,

$$\begin{aligned} \eta\nu c_n[1 - \Psi_{Z_1}(tc_n^{-1/2})e^{-itc_n^{-1/2}E[Z_1(v_n)]}e^{isc_n^{-1/2}} + isc_n^{-1/2}] \\ = \eta\nu c_n[(1 - \Psi_{Z_1}(tc_n^{-1/2})e^{-itc_n^{-1/2}E[Z_1(v_n)]}) + (1 - e^{isc_n^{-1/2}} + isc_n^{-1/2}) + o(c_n^{-1})] \\ \xrightarrow[n \rightarrow +\infty]{} \frac{\eta\nu(\sigma^2 - 1/\eta^2)t^2}{2} + \frac{\eta\nu s^2}{2}. \end{aligned}$$

This along with (4.5) implies that

$$c_n^{-1/2} \begin{bmatrix} \overline{N}_n^* - \hat{N}_n^*E[Z_1(v_n)] \\ \hat{N}_n^* - E[\hat{N}_n^*] \end{bmatrix} \xrightarrow[n \rightarrow +\infty]{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \eta\nu(\sigma^2 - \frac{1}{\eta^2}) & 0 \\ 0 & \nu\eta \end{bmatrix}\right),$$

which together with (4.4) and the fact that  $E[Z_1(v_n)] \xrightarrow[n \rightarrow +\infty]{} 1/\eta$  yields the result.  $\square$

The asymptotic normality of  $\hat{\eta}_n^*$  is now an immediate consequence of the previous results.

**Corollary 4.1** *If the conditions of Theorems 4.1 and 4.2 hold then*

$$\sqrt{c_n}(\hat{\eta}_n^* - \eta_n) \xrightarrow[n \rightarrow +\infty]{d} \mathcal{N}\left(0, \sqrt{\frac{\eta}{\nu}(\eta^2\sigma^2 - 1)}\right),$$

where  $\eta_n = E[\hat{N}_n^*]/E[\overline{N}_n^*]$ .

**Proof:** Conditions (4.1)-(4.3) imply the conditions of Theorems 3.1 and 3.2, hence  $c_n^{-1}\bar{N}_n \xrightarrow[n \rightarrow +\infty]{P} \nu$ . On the other hand condition  $\tilde{D}^{(3)}(v_n)$  implies that  $\eta_n \xrightarrow[n \rightarrow +\infty]{} \eta$ . The result now follows from the fact that

$$\sqrt{c_n}(\hat{\eta}_n - \eta_n) = \frac{1}{c_n^{-1}\bar{N}_n} \{c_n^{-1/2}(\hat{N}_n - E[\hat{N}_n]) - \eta_n c_n^{-1/2}(\bar{N}_n - E[\bar{N}_n])\}.$$

□

**Remark 8** Since the variance of  $\hat{\eta}_n^*$  is of order  $1/\nu$ , if we kept  $\nu$  fixed as in (1.2) we could not guarantee the consistency of  $\hat{\eta}_n^*$ . Hence the need to assume that  $\nu = \nu_n \rightarrow +\infty$  as  $n \rightarrow +\infty$ .

To conclude we present a result that enables one to construct approximate confidence intervals or a hypothesis test regarding  $\eta$ , i.e., to determine the extent of clustering of upcrossings of high levels in the observed data. From a practical viewpoint it is more useful than Corollary 4.1.

**Corollary 4.2** Suppose that the conditions of Corollary 4.1 hold,

$$\sqrt{c_n}(\eta_n - \eta) \xrightarrow[n \rightarrow +\infty]{} 0$$

and

$$c_n^{-1}E[(Z_1^*(v_n))^4 \mathbb{I}_{\{(Z_1^*(v_n))^2 \leq c_n\}}] \xrightarrow[n \rightarrow +\infty]{} 0.$$

Then

$$\sqrt{\frac{\bar{N}_n}{\hat{\eta}_n^*((\hat{\eta}_n^* \hat{\sigma}_n)^2 - 1)}} (\hat{\eta}_n^* - \eta) \xrightarrow[n \rightarrow +\infty]{} \mathcal{N}(0, 1),$$

$$\text{where } \hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n-3} Y_i^2 \mathbb{I}_{\{\tilde{N}_{i,i}=0, X_{i+2} \leq v_n < X_{i+3}\}}}{\hat{N}_n}.$$

**Proof:** Straightforward from Corollary 4.1 and the fact that Theorems 3.1 and 3.2 imply that

$$c_n^{-1} \sum_{i=1}^{n-3} Y_i^2 \mathbb{I}_{\{\tilde{N}_{i,i}=0, X_{i+2} \leq v_n < X_{i+3}\}} \xrightarrow[n \rightarrow +\infty]{P} \eta \nu \sigma^2.$$

□

We recall that the properties proved for the estimator  $\hat{\eta}_n^*$  remain valid for the estimator  $\hat{\eta}_n$  used in the sequel of this paper.

## 5 The choice of the levels

In Section 3 it was shown that for the runs estimator of the upcrossings index  $\hat{\eta}_n$  consistency can be achieved by the use of somewhat lower levels  $v_n$  satisfying (3.1). Thus, the precise choice of  $v_n$  depends on the knowledge of the joint distribution of  $(X_1, X_2)$ , typically unknown. These deterministic levels will have to be replaced, in practical situations, by random levels suggested by the relation

$$\frac{n}{c_n} P(X_1 \leq v_n < X_2) \sim \nu. \quad (5.6)$$

This relation basically states that the expected number of upcrossings of the level  $v_n$  is approximately  $c_n \nu$ . Contrarily to the random levels used in the estimation of the extremal index, these random levels can not be represented by an appropriate order statistic. Nevertheless, for levels  $v_n$  such that  $\frac{n}{c_n} P(X_1 > v_n) \sim \tau$  the expected number of upcrossings will be necessarily smaller or equal to  $c_n \tau$ . It then seems natural to also replace  $v_n$  by the appropriate order statistic used in the estimation of the extremal index (for example, Nandagopalan [15]), namely  $\hat{v}_n = X_{n-[c_n \tau]:n}$ . With this consideration we obtain the following estimator for  $\eta$

$$\tilde{\eta}_n = \frac{\hat{N}_n(\hat{v}_n)}{\tilde{N}_n(\hat{v}_n)},$$

where  $\hat{N}_n(\hat{v}_n) = \hat{N}_n([0, 1])$  and  $\tilde{N}_n(\hat{v}_n) = \tilde{N}_n([0, 1])$  with  $u_n = \hat{v}_n$ .

The weak consistency of this estimator can be obtained by showing that it is closely approximated by the corresponding estimates based on the non-random levels  $v_n$ . For this, let's start by noting that for two levels  $v_1$  and  $v_2$ , such that  $v_1 \neq v_2$ , and fixed  $n$  we have

$$|\tilde{N}_n(v_1) - \tilde{N}_n(v_2)| \leq |N_n(v_1) - N_n(v_2)| \quad (5.7)$$

and

$$|\hat{N}_n(v_1) - \hat{N}_n(v_2)| \leq |N_n(v_1) - N_n(v_2)| \quad (5.8)$$

where  $N_n(v_n) = \sum_{i=1}^n \mathbb{I}_{\{X_i > v_n\}}$ .

**Theorem 5.1** *Suppose that for each  $\nu > 0$  there exists  $u_{[n/c_n]}^{(\nu)} = u_{[n/c_n]}^{(\tau)}$  for some  $\tau > 0$ , the conditions of Proposition 5.3.1 and 5.3.2 in Nandagopalan [15] hold for each  $\tau'$  in a neighbourhood of  $\tau$  and conditions (3.3), (3.7) and (3.8) hold for each  $\nu'$  in a neighbourhood of  $\nu$ . Then  $\tilde{\eta}_n \xrightarrow[n \rightarrow +\infty]{P} \eta$ .*

**Proof:** For  $\epsilon > 0$

$$\begin{aligned} & P(|c_n^{-1} \tilde{N}_n(\hat{v}_n) - \nu| > 6\epsilon) \\ & \leq P(|c_n^{-1} [c_n \tau] - N_n(v_n)| > 3\epsilon) + P(|c_n^{-1} \tilde{N}_n(v_n) - \nu| > 3\epsilon) \end{aligned} \quad (5.9)$$

by (5.7), with  $v_1$  and  $v_2$  replaced by  $\hat{v}_n$  and  $v_n = u_{[n/c_n]}(\tau + \epsilon)$ . If  $\epsilon$  is sufficiently small, then it follows from the results in [15] and in Section 3 that  $c_n^{-1} N_n(v_n) \xrightarrow[n \rightarrow +\infty]{P} \tau + \epsilon$  and  $c_n^{-1} \tilde{N}_n(v_n) \xrightarrow[n \rightarrow +\infty]{P} \nu + \epsilon$ . Therefore, since  $c_n^{-1} [c_n \tau] \xrightarrow[n \rightarrow +\infty]{} \tau$ , for large  $n$  (5.9) is dominated by

$$P(|c_n^{-1} N_n(v_n) - (\tau + \epsilon)| > \epsilon) + P(|c_n^{-1} \tilde{N}_n(v_n) - (\nu + \epsilon)| > \epsilon) \xrightarrow[n \rightarrow +\infty]{} 0,$$

which proves that  $c_n^{-1} \tilde{N}_n(\hat{v}_n) \xrightarrow[n \rightarrow +\infty]{P} \nu$ . On the other hand, for  $\epsilon > 0$ , we have from (5.8), with  $v_1$  and  $v_2$  replaced by  $\hat{v}_n$  and  $v_n = u_{[n/c_n]}(\tau + \epsilon)$ , that

$$\begin{aligned} & P(|c_n^{-1} \hat{N}_n(\hat{v}_n) - \eta \nu| > 6\epsilon) \\ & \leq P(|c_n^{-1} [c_n \tau] - N_n(v_n)| > 3\epsilon) + P(|c_n^{-1} \hat{N}_n(v_n) - \nu| > 3\epsilon). \end{aligned} \quad (5.10)$$

Hence, if  $\epsilon$  is sufficiently small we have  $c_n^{-1} \hat{N}_n(v_n) \xrightarrow[n \rightarrow +\infty]{P} \eta(\nu + \epsilon)$  (results of Section 3) and can may conclude that (5.10) is dominated by

$$P(|c_n^{-1} N_n(v_n) - (\tau + \epsilon)| > \epsilon) + P(|c_n^{-1} \hat{N}_n(v_n) - \eta(\nu + \epsilon)| > \epsilon) \xrightarrow[n \rightarrow +\infty]{} 0,$$

which proves that  $c_n^{-1} \widehat{N}_n(\widehat{v}_n) \xrightarrow[n \rightarrow +\infty]{P} \eta\nu$ .

The result is now an immediate consequence of the two convergences established.  $\square$

The proof of the asymptotic normality of  $\widetilde{\eta}_n$  remains an open problem.

## 6 Simulations: Some examples

This section studies some examples of sequences exhibiting clustering of upcrossings. We illustrate the performance of the runs estimator of the upcrossings index with the max-autoregressive (ARMAX) process of Ferreira [4] and the negatively correlated first order autoregressive (AR(1)) process of Chernick *et al.* [2], for which the values of the upcrossings index  $\eta$  are well known.

### 6.1 ARMAX process

Lets consider the ARMAX process of Ferreira [4],

$$X_n = \max\{Y_n, Y_{n-2}, Y_{n-3}\}, \quad n \geq 1, \quad (6.1)$$

where  $\{Y_n\}_{n \geq -2}$  is a sequence of independent and uniformly distributed on  $[0, 1]$  variables. In Figure 1 we present a sample path of this process.

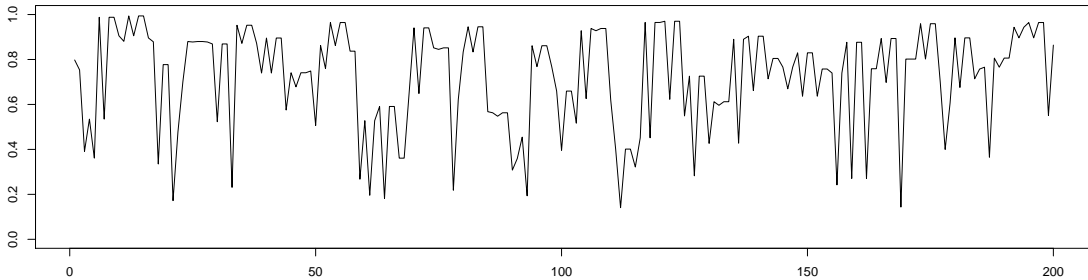


Figure 1: Sample path of the stationary ARMAX process in (6.1).

Condition  $\widetilde{D}^{(3)}(\mathbf{u})$  holds for this stationary sequence, with  $\mathbf{u} = \mathbf{u}^{(\nu)} = \{u_n = 1 - \tau/n\}_{n \geq 1}$ ,  $\tau > 0$ . It has extremal index  $\theta = 1/3$  and upcrossings index  $\eta = 1/2$ . Furthermore, it holds  $\lim_{n \rightarrow +\infty} nP(X_1 > u_n) = 3\tau$  and  $\lim_{n \rightarrow +\infty} nP(X_1 \leq u_n < X_2) = 2\tau$ , so the levels  $u_n$  are simultaneously normalized for exceedances and upcrossings. Note that this implies that for a sample  $(X_1, \dots, X_n)$  of (6.1) with  $n$  sufficiently large, the number of upcrossings of a high level is approximately  $2/3$  of the number of exceedances of the same level.

We shall consider in the following simulations the level  $u = X_{n-k:n}$ , in (2.3) corresponding to the  $(k+1)$ th top order statistics associated to the random sample  $(X_1, \dots, X_n)$  of (6.1), commonly used in the estimation of the extremal index (see, for example, Nandagopalan [15] and Gomes, *et al.* [7]). Hence, the upcrossings index estimator is now a function of  $k$ ,  $\widehat{\eta}_n(k)$ .

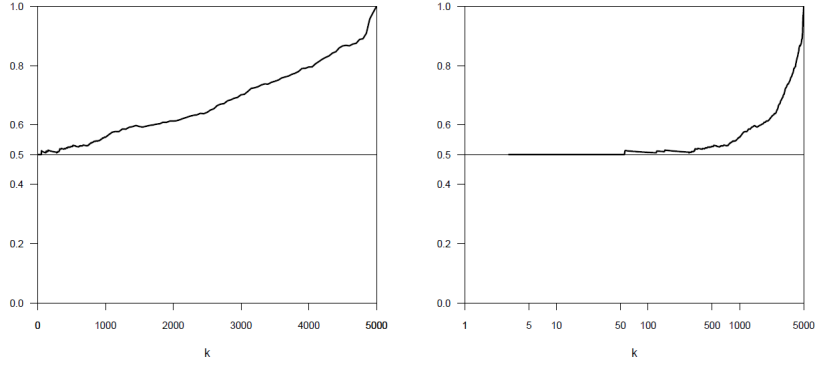


Figure 2: Sample path of the upcrossings index estimator as a function of  $k$  for a sample of size  $n = 5000$  from the stationary ARMAX process in (6.1), in a linear scale (left) and in a logarithmic scale (right).

In Figure 2 we present a sample path of the estimator  $\hat{\eta}_n$  in (2.3) as a function of  $k \geq 1$ , for a sample of size  $n = 5000$  in a linear scale and in a logarithmic scale.

As we can see from Figure 2 the logarithmic scale enhances the performance of  $\hat{\eta}_n$  for small values of  $k$ , giving better insight of its stability region around  $\eta = 0.5$ .

For samples of size  $n = 100, 200, 500, 1000, 2000, 5000$  and  $10000$ , from the ARMAX process in (6.1), we have performed a multi-sample Monte Carlo simulation with 5000 runs and 10 replicates. For details on multi-sample simulation see, for instance, Gomes and Oliveira [8]. We have simulated for the estimator  $\hat{\eta}_n(k)$  in (2.3), the mean value ( $E$ ), the mean squared error ( $MSE$ ) and the optimal sample fraction  $k_0$  with  $k_0 := \arg \min_k MSE[\hat{\eta}_n(k)]$ .

In Figure 3, we illustrate the values of the estimated mean values and MSE's of  $\hat{\eta}_n(k)$  in (2.3), for a sample of size  $n = 5000$  from the ARMAX process in (6.1), with upcrossings index  $\eta = 0.5$ . In Table 1 we present the main distributional properties of the estimator under study with the associated 95% confidence intervals (see Gomes and Oliveira [8]).

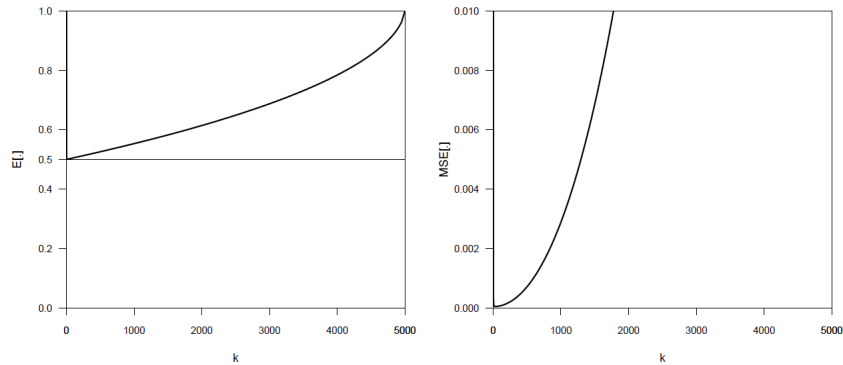


Figure 3: Estimated mean values (left) and mean squared errors (right), for samples of size  $n = 5000$  from the ARMAX process in (6.1) ( $\eta = 0.5$ )

$n$	$k_0$	$k_0/n$	$E[\bullet]$	$MSE[\bullet]$	$SD[\bullet]$
100	16	0.160	$0.52735 \pm 0.00042$	$0.00497 \pm 0.000087$	$0.06495 \pm 0.00064$
200	19	0.095	$0.51754 \pm 0.00029$	$0.00214 \pm 0.000029$	$0.04282 \pm 0.00037$
500	25	0.050	$0.50977 \pm 0.00021$	$0.00072 \pm 0.000018$	$0.02493 \pm 0.00030$
1000	30	0.030	$0.50622 \pm 0.00014$	$0.00033 \pm 0.000008$	$0.01699 \pm 0.00022$
2000	36	0.018	$0.50389 \pm 0.00007$	$0.00015 \pm 0.000003$	$0.01162 \pm 0.00013$
5000	54	0.011	$0.50241 \pm 0.00006$	$0.00005 \pm 0.000001$	$0.00694 \pm 0.00010$
10000	67	0.007	$0.50152 \pm 0.00003$	$0.00003 \pm 0.000001$	$0.00477 \pm 0.00006$

Table 1: Optimal sample fractions, mean values and mean squared errors of the estimator at its optimal levels, for the ARMAX process, with  $\eta = 0.5$ .

## 6.2 AR(1) process

Lets consider the negatively correlated uniform AR(1) process of Chernick *et al.* [2],

$$X_n = -\frac{1}{r}X_{n-1} + \epsilon_n, \quad n \geq 1, \quad (6.2)$$

where  $\{\epsilon_n\}_{n \geq 1}$  is a sequence of independent and identically distributed random variables, such that, for a fixed integer  $r \geq 2$ ,  $\epsilon_n \sim U\{\frac{1}{r}, \dots, \frac{r-1}{r}, 1\}$  and  $X_0 \sim U(0, 1)$  independent of  $\epsilon_n$ .

Condition  $\tilde{D}^{(3)}(\mathbf{u})$  also holds for this stationary sequence, with  $\mathbf{u} = \mathbf{u}^{(\nu)} = \{u_n = 1 - \tau/n\}_{n \geq 1}$ ,  $\tau > 0$ , moreover  $\lim_{n \rightarrow +\infty} nP(X_1 > u_n) = \tau$  and  $\eta = \theta = 1 - 1/r^2$  (see Sebastião *et al.* [18]). Condition  $D''(\mathbf{u})$  typically doesn't hold for these sequences since they tend to oscillate rapidly near extremes. To illustrate this characteristic we present in Figure 4 sample paths of the negatively correlated uniform AR(1) processes for  $r = 2$ ,  $r = 3$  and  $r = 5$  ( $\eta = \theta = 0.75$ ,  $\eta = \theta = 0.89$  and  $\eta = \theta = 0.96$ , respectively).

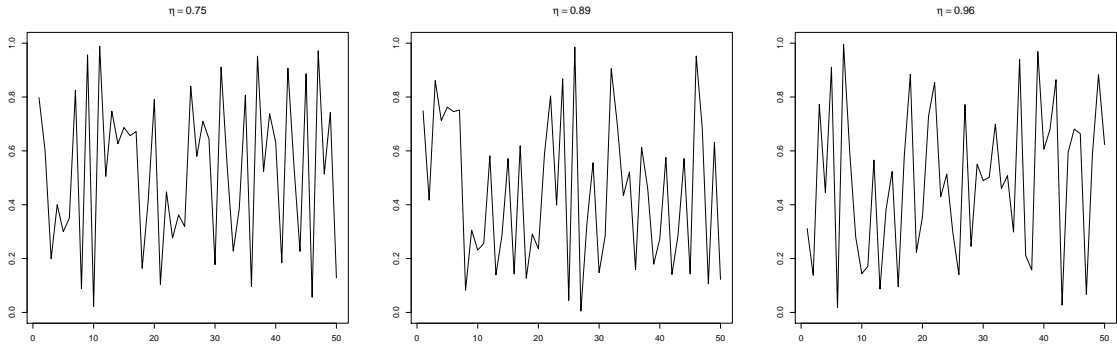


Figure 4: Sample paths of the stationary AR(1) process in (6.2), with  $r = 2$  (left),  $r = 3$  (center) and  $r = 5$  (right)

We present, in Figures 5-7, the estimated mean values and MSE's of  $\hat{\eta}_n(k)$  in (2.3), for a sample of size  $n = 5000$  from the AR(1) process in (6.2), with upcrossings index  $\eta = 0.75$ ,  $0.89$  and  $0.96$ , respectively.

In Table 2 we present the main distributional properties of the estimator under study with the associated 95% confidence intervals.

## 6.3 Some overall conclusions

- The sample paths of the runs estimator of the upcrossings index have very different patterns for the ARMAX and the AR(1) process. Whilst for the ARMAX process the estimates increase with

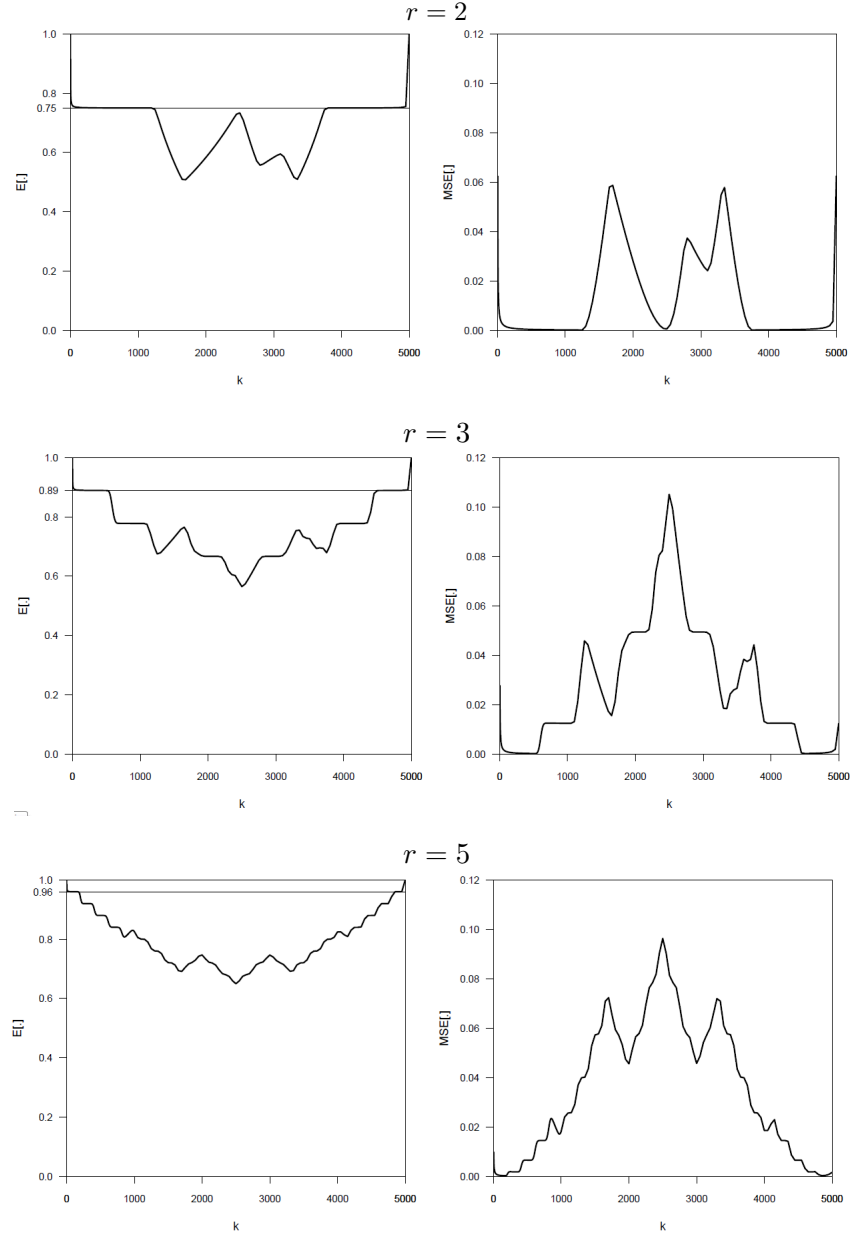


Figure 5: Estimated mean values (left) and mean squared errors (right), for samples of size  $n = 5000$  from the AR(1) process in (6.2) with, respectively,  $r = 2$  ( $\eta = 0.75$ ),  $r = 3$  ( $\eta = 0.89$ ) and  $r = 5$  ( $\eta = 0.96$ )

the value of  $k$ , for the AR(1) process the estimates, as function of  $k$ , have almost a symmetric distribution, decreasing with smaller values of  $k$ .

- For the ARMAX process, smaller values of  $\eta$  tend to be associated with a wider “bathtub” pattern of the mean squared error as a function of  $k$ .
- The runs estimator of the upcrossings index has some mean value stability around the target  $\eta$  for small values of  $k$ , exhibiting, for such values, the mean squared error a “bathtub” pattern, although not very wide.
- For small values of  $k$ , or equivalently high levels, we obtain good estimates for  $\eta$ . We remark that



$n$	$k_0$	$k_0/n$	$E[\bullet]$	$MSE[\bullet]$
<b><math>\eta = 0.75</math></b>				
100	24	0.240	$0.72839 \pm 0.00081$	$0.00769 \pm 0.000075$
200	49	0.245	$0.73228 \pm 0.00043$	$0.00364 \pm 0.000055$
500	123	0.246	$0.73942 \pm 0.00026$	$0.00140 \pm 0.000016$
1000	246	0.246	$0.74375 \pm 0.00009$	$0.00068 \pm 0.000009$
2000	494	0.247	$0.74576 \pm 0.00017$	$0.00034 \pm 0.000005$
5000	1240	0.248	$0.74745 \pm 0.00010$	$0.00013 \pm 0.000002$
<b><math>\eta = 0.89</math></b>				
100	10	0.100	$0.86547 \pm 0.00112$	$0.01254 \pm 0.000240$
200	19	0.095	$0.87623 \pm 0.00077$	$0.00598 \pm 0.000096$
500	48	0.096	$0.88347 \pm 0.00029$	$0.00221 \pm 0.000025$
1000	98	0.098	$0.88642 \pm 0.00017$	$0.00105 \pm 0.000009$
2000	201	0.101	$0.88768 \pm 0.00020$	$0.00050 \pm 0.000007$
5000	519	0.104	$0.88842 \pm 0.00007$	$0.00019 \pm 0.000003$
<b><math>\eta = 0.96</math></b>				
100	7	0.070	$0.91442 \pm 0.00123$	$0.01431 \pm 0.000366$
200	9	0.045	$0.93962 \pm 0.00081$	$0.00728 \pm 0.000126$
500	17	0.034	$0.95373 \pm 0.00072$	$0.00274 \pm 0.000077$
1000	33	0.033	$0.95687 \pm 0.00054$	$0.00130 \pm 0.000038$
2000	67	0.034	$0.95842 \pm 0.00026$	$0.00061 \pm 0.000011$
5000	173	0.035	$0.95949 \pm 0.00017$	$0.00023 \pm 0.000004$

Table 2: Optimal sample fractions, mean values and mean squared errors of the estimator at its optimal levels, for the AR(1) process, with  $\eta = 0.75$ ,  $\eta = 0.89$  and  $\eta = 0.96$ .

the choice of the number of top order statistics is a complex problem in extreme value applications.

- The sensitivity of the mean value to the changes in  $k$  seem to clarify the need of studying the bias properties of this estimator. This is clear from Figure 6 where we can see that for the ARMAX process the behaviour of the mean squared error is almost determined by the bias (overlap of both curves), since the variance is always very small. The same behaviour holds for the AR(1) process.

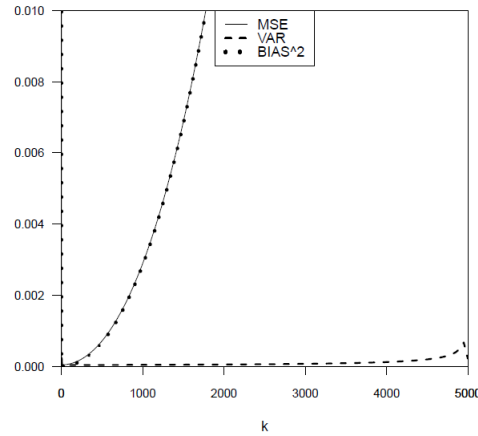


Figure 6: Estimated mean squared error (solid line), estimated variance (dashed line) and estimated bias (dotted line), for samples of size  $n = 5000$  from the ARMAX process in (6.1).

## 7 Case-studies

### 7.1 Ozone pollution

We now consider the performance of the above mentioned estimator in the analysis of  $n = 120$  weekly maxima of hourly averages of ozone concentrations measured in parts per million, in the San Francisco bay area, San José. These data are available in the package Xtremes (Reiss and Thomas [16]) and have already been studied, for instance, in Gomes *et al.* [7] when estimating the extremal index. In Figure 7 we picture the data over the above mentioned period.

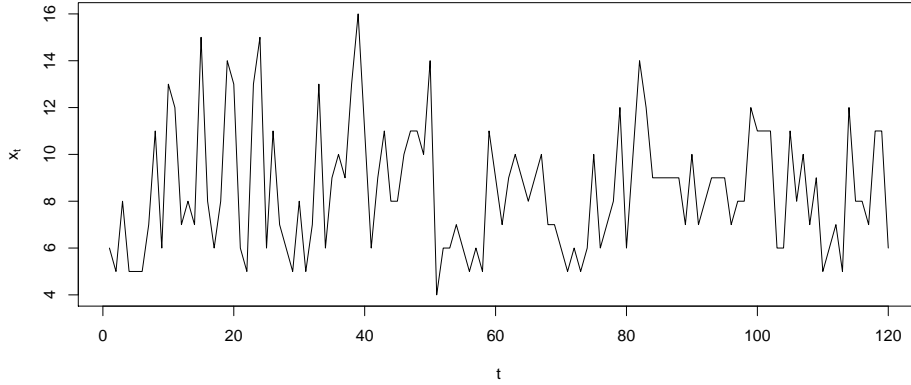


Figure 7: Weekly maxima of hourly averages of ozone concentrations measured in parts per million, in the San Francisco bay area, San José

Since in practice condition  $\tilde{D}^{(3)}$  is not yet possible to verify we assumed that it holds since as stated Gomes *et al.* [7], most of the parametric models that adequately fit this type of meteorological data satisfy condition  $D''$  and therefore also satisfy condition  $\tilde{D}^{(3)}$ .

In Figure 8 we picture the sample paths of  $\hat{\eta}_n(k)$  as functions of  $k$  in a linear scale (left) and logarithmic (right) scale.

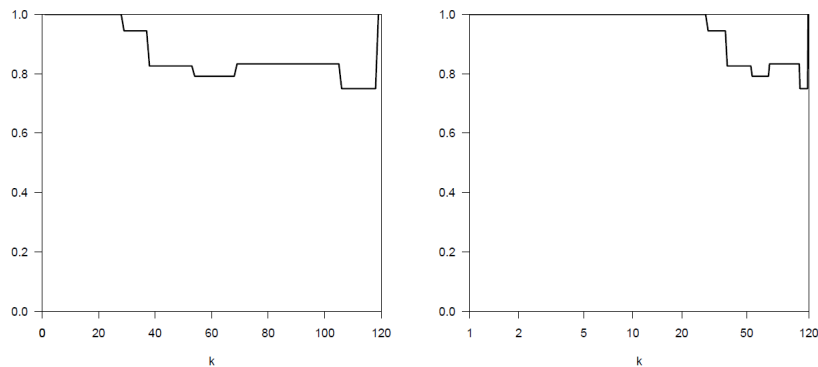


Figure 8: Estimates of the upcrossings index as a function of  $k$  for the weekly maxima of hourly averages of ozone concentrations measured in parts per million, in the San Francisco bay area, San José, in a linear scale (left) and in a logarithmic scale (right).

The stability around one for small values of  $k$  agrees with the fact that  $\eta = 1$  since from [7] condition  $D''$  holds. Nevertheless, the small number of observations makes it difficult to rely on such a point estimate since the number of upcrossings and consequently the number of non-upcrossings followed by an upcrossing, of a high level, will always be very small.

## 7.2 Financial log-returns

Financial time series are very unlikely to satisfy condition  $D''$  because of their varying volatility. Therefore, we shall consider now the time series of log returns of the German stock market index DAX, consisting of the 30 major German companies trading on the Frankfurt Stock Exchange. The daily closing prices of the DAX index for the period from 1991 to 1998 that we work with is available as dataset `EuStockMarkets` in the statistics software R, which we also used for all computations. This data set of 1786 observations has been considered in Klar *at al.* [10] where we can find a full statistical description. In Figure 9 we picture DAX daily closing prices over the mentioned period,  $x_t$ , and the log-returns,  $100 \times (\ln x_t - \ln x_{t-1})$ , the data to be analyzed.

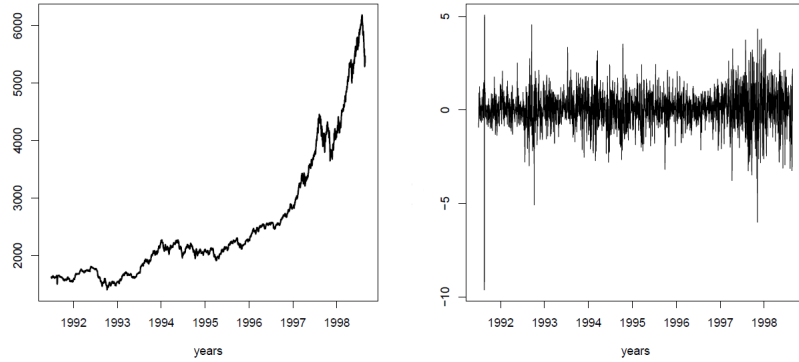


Figure 9: DAX daily closing prices (left) and daily log-returns (right)

This time series of log returns is well modeled by a GARCH(1,1), covariance stationary, process ([10]). Using a goodness-of-fit test Klar *at al.* [10], at a 5% level, do not reject the hypothesis that the innovations of the process follow a t-distribution with seven degrees of freedom.

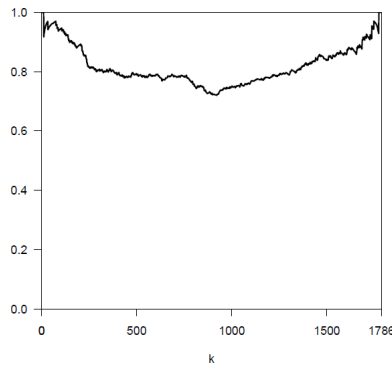


Figure 10: Estimates of the upcrossings index as a function of  $k$  for the DAX log-returns

The graph in Figure 10 allows us to conclude that  $\eta$  is not equal to one, which agrees with the fact that these processes are very unlikely to verify condition  $D''$ . The sample path exhibits a stability region

around a value close to  $\eta = 0.8$ .

Mikosch and Stărică [14] derive the extremal index for the squared GARCH(1,1) processes, whereas Laurini and Tawn [11] propose an algorithm for the evaluation of the extremal index of GARCH(1,1) processes with  $t$ -distributed innovations. It would be interesting, in future work, to obtain similar results for the upcrossings index of a GARCH(1,1) process.

## 8 Conclusions

The upcrossings index, as a measure of the clustering of upcrossings of high levels, is an important parameter when studying extreme events. For sequences satisfying condition  $\tilde{D}^{(3)}$ , that locally restricts the dependence of the sequence but still allows clustering of upcrossings, we have proposed a simple estimator for this parameter. The study of the properties of the proposed estimator, namely its consistency and asymptotic normality, has been carried out in Sections 3 and 4. With simulations of well known autoregressive processes that verify condition  $\tilde{D}^{(3)}$  we were able to illustrate the performance of the estimator. Case studies in the fields of environment and finance were also exploited.

Relation (1.5) allows one also to estimate  $\eta$  through the extremal index  $\theta$  modified by consistent estimators of exceedances,  $\tau$ , and the mean number of upcrossings,  $\nu$ , of high levels. Several estimators for the extremal index can be found in the literature (see Ancona-Navarrete and Tawn [1] for a survey). Other estimators arise from Proposition 1.1 and from the relation with  $\mu_1$ , the upcrossings-tail dependence coefficient. In future work we intend to propose new estimators for the upcrossings index and compare the several estimating methods.

## References

- [1] Ancona-Navarrete, M. A. and Tawn, J. A. (2000). A comparison of methods for estimating the extremal index. *Extremes*, **3**(1), 5-38.
- [2] Chernick, M., Hsing, T. and McCormick, W. (1991). Calculating the extremal index for a class of stationary sequences. *Adv. Appl. Prob.*, **23**, 835-850.
- [3] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. New York: Wiley.
- [4] Ferreira, H. (2006). The upcrossing index and the extremal index. *J. Appl. Prob.*, **43**, 927-937.
- [5] Ferreira, H. (2007). Runs of high values and the upcrossings index for a stationary sequence. In *Proceedings of the 56th Session of the ISI*.
- [6] Ferreira, M. and Ferreira, H. (2011). On extremal dependence: some contributions. *Test*. DOI: 10.1007/s11749-011-0261-3.
- [7] Gomes, M.I., Hall, A. and Miranda, C. (2008). Subsampling techniques and the Jackknife methodology in the estimation of the extremal index. *J. Statist. Comput. Simulation*, **52**, 2022-2041.
- [8] Gomes, M.I. and Oliveira, O. (2001). The bootstrap methodology in Statistics of extremes - choice of the optimal sample fraction. *Extremes*, **4**(4), 331-358.

- [9] Hsing, T., Hüsler, J. and Leadbetter, M.R. (1988). On the exceedance point process for a stationary sequence. *Prob. Th. Rel. Fields*, 78, 97-112.
- [10] Klar, B., Lindner, F. and Meintanis, S.G. (2011). Specification tests for the error distribution in Garch models. *Computational Statistics and Data Analysis*. DOI:10.1016/j.csda.2010.05.029
- [11] Laurini, F. and Tawn, J. (2006). The extremal index for GARCH(1,1) processes with t-distributed innovations. <http://www.scientificcommons.org/17368349>
- [12] Leadbetter, M.R. (1974). On extreme values in stationary sequences. *Z. Wahrsch. Verw. Gebiete*, 28, 289-303.
- [13] Leadbetter, M. R. and Nandagopalan, S. (1988). On exceedance point process for stationary sequences under mild oscillation restrictions. In J. Hüsler and D. Reiss (eds.). *Extreme Value Theory: Proceedings, Oberwolfach 1987*, Springer, New York, 69-80.
- [14] Mikosch, T. and Stărică, C. (2000). Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Ann. Statist.*, **28**, 1427-1451.
- [15] Nandagopalan, S. (1990). *Multivariate Extremes and Estimation of the Extremal Index*. Ph.D. Thesis, University of North Carolina. Chapel Hill.
- [16] Reiss, R.-D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Verlag, Basel, Switzerland.
- [17] Ross, S. (1983). *Stochastic Processes*. New York: Wiley.
- [18] Sebastião, J., Martins, A.P., Pereira, L. and Ferreira, H., (2010). Clustering of upcrossings of high values. *J. Statist. Plann. Inf.*, **140**, 1003–1012